II. RESOURCE OPERATIONS

A. DESCRIPTION OF PROGRESS

B. SUMMARY OF RESOURCE USAGE

C. RESOURCE RELATED RESEARCH EQUIPMENT LIST

D. SUMMARY OF PUBLICATIONS

A. DESCRIPTION OF PROGRESS

OVERVIEW

In the first twelve months of this fifteen-month grant period, the DENDRAL programs and the GC/MS data system have moved significantly forward under NIH funding, even though it was partly a time of transition from one computer system to another. This report of progress is organized in three parts, corresponding to the three specific aims of our December, 1973, proposal: (PART 1) Enhancing the power of the mass spectrometry resource, (PART 2) Developing performance and theory formation programs, and (PART 3) Applying the computer programs and instrumentation to biomedically relevant structure elucidation problems.

The highlight of the period since May 1, 1974, was the project's move to the interactive computing environment of the NIH-funded SUMEX-AIM facility from the batch computing environment of the Stanford Computation Center. Because of this, many scientists outside this university have been able to use the DENDRAL computer programs for their own research. Also, the programs themselves grew in power and scope, and we opened new vistas for collaboration with other research groups. We have been able to make the programs more conversational and thus more helpful to the chemists and biochemists for whom they were developed. Outside users in other research groups also have in SUMEX an easy mechanism for trying out the DENDRAL programs on their own structure elucidation problems. Finally, we have a mechanism for looking at subroutines developed by other research groups in the context of our own programs -- and have incorporated subroutines written, for example, by T. Wipke and by R. Feldmann, into our procedures. The programs and their development are discussed in Part 2, below.

The DENDRAL project, one of the major users of the SUMEX-AIM computer facility, has been forming its own community of remote users. This "exodendral" community has already provided valuable contributions to program development and both the community and contributions are expected to grow at an increased rate. As an example, for the last month for which figures are available (March 1975), the number of CPU hours used by exodendral persons amounted to at least ten percent of the CPU hours used by the entire DENDRAL project. In the last month alone, one new exodendral account representing at least three users has been added to the system, and another four exodendral users have been invited to begin their usage via various "guest" accounts.

Another milestone in this period was the delivery of the PDP-11/45 computer, and successful transfer of data acquisition and reduction programs into that computer. This has provided a stand-alone environment for our mass spectrometer/computer system

in which development, experimentation and routine use of the system is much more simple, reliable and efficient than previously. We have made excellent progress in fulfilling the goals of combined gas chromatography/high resolution mass spectrometry (see Part 1, below).

Our programs are receiving heavy use from local users and outside users who are investigating mass spectrometry problems for a variety of different compound classes. In addition, new program developments have extended the scope of biomedical structure elucidation problems for which we can provide some computer assistance. Local users include members of Professor Djerassi's group, other chemistry department persons and research groups at the Stanford Medical School. We have recently begun the process of building a community of outside users who can access our programs at SUMEX via TYMNET or ARPANET. Several research groups have expressed considerable interest; we have demonstrated and explained the programs to several groups and we are currently arranging more demonstrations and assisting other people in learning to use SUMEX and the programs from their own laboratories. These applications are discussed in detail in Part 3, below.

1 PART 1: ENHANCING THE POWER OF THE M.S. RESOURCE

1.1 Introduction

Our grant proposal requested funds for significant upgrading of our capabilities in mass spectrometry. The goals of this upgrading were to provide routine high resolution mass spectrometry (HRMS), combined gas chromatography/low resolution mass spectrometry (GC/LRMS) and to develop a combined gas chromatography/high resolution mass spectrometry (GC/HRMS) facility. In addition, this would provide the capability for new experiments in the detection and utilization of data on metastable ions. These capabilities would then be available as required for application to our wider goal, solution of biomedical structure elucidation problems of community of researchers.

The upgrading included several items of hardware and software development, as follows: 1) Acquire stand-alone computer support for the mass spectrometer because existing facilities were inadequate and very expensive; 2) convert existing software, written in the PL/ACME language into FORTRAN so that it would run on the new system; 3) develop new software as required for the demanding task of GC/HRMS; 4) provide hardware and software for semi-automatic acquisition of data on metastable ions. The initial development phase of this upgrading included performance tests to determine the capabilities and limitations of the GC/HRMS system to define the scope of problems to which it can be applied.

The present status of this effort is that the computer system has been purchased, installed and is operating. The software has been converted and is operational. The GC/HRMS system is in the trial stage and is working. Future developments include significant improvements in the software to provide more routine and reliable GC/HRMS operation and to provide better information to the operator on instrument performance and to the chemist on the characteristics of his data. The metastable ion work has been deferred until now because of the more pressing demands of the GC/HRMS system, but work can now begin on this aspect of our research.

We presently are in a position to provide routine LRMS and HRMS support for our chemical research and program development. The GC/HRMS system is working well enough to commence study of real problems. The above developments and future goals are summarized in detail in the subsequent sections.

1.2 Hardware Acquisition and Development

We have, in the mass spectrometry laboratory, two mass spectrometers which were connected to our previous computer system (ACME), the Varian-MAT 711 mass spectrometer, and the AEI MS-9, both high resolution mass spectrometers. We have concentrated our efforts to this point on development of the 711/computer system because this (much more modern) instrument is the spectrometer of choice for the GC/HRMS experiments. It was already equipped with the gas chromatograph and GC/LRMS work was already routine as far as the mass spectrometer system was concerned. We were granted some money for minor upgrading of the MS-9 so that it could relieve some of the burden of routine HRMS analysis from the 711 (see Future Developments).

At the time the grant was awarded, we were essentially without computer support in the mass spectrometry laboratory. Interim funds were provided to permit us to connect to the ACME system running on a different computer system. This connection was made and permitted us to obtain some HRMS data while purchase and installation of the stand-alone system was completed. The interim ACME system was even less effective than in its previous environment, and did not permit GC/MS experiments due to slow response time.

We concurred with the study section's recommendation that stand-alone computer support be provided for efficiency and long-term cost effectiveness, and that such support be a PDP 11/45 or equivalent as the machine with the capabilities to handle the heavy data burden imposed by GC/HRMS. We were able to adjust our first year budget to allow purchase of this computer. It was ordered on Feb. 11, 1974 as a contingent order (contingent on award of the grant). The firm order was placed March 18, 1974. It was actually delivered on August 2, 1974. Together with the Digital Equipment Corp. (DEC) PDP-11/45, we obtained a disk system from Systems Industries because it was considerably less expensive than the comparable DEC device.

A diagram of the current hardware system is shown in Figure 1. Note that this system is interfaced to the mass spectrometer through a previously existing PDP-11/20. There are two important reasons for this configuration. The 11/20 contained the necessary hardware extensions for computer control of the mass spectrometer under the old ACME system. This drastically reduced the mass spectrometer/computer interface problems. The 11/20 acts as a buffer between the mass spectrometer and the 11/45, thus freeing the 11/45 for computations during the course of data acquisition. This is an important element of future foreground/background processing.

1.3 Software Development

Conversion of existing PL/ACME programs to FORTRAN was begun on the award of the grant. The delays in delivery of the 11/45 system caused delays in this development because no machine was available for certain tests of the programs, and of course, no work on new mass spectral data could be done until the mass spectrometer and computer system became operational. Conversion of these algorithms also included many system software developments to ensure that previously batch processing programs could function in a real-time environment under the requirements of GC/HRMS operation. This development included not only improvements and extensions to existing algorithms, but building a file management system for facile logging and storage of spectra with the ability for simple recall to examine or recompute old data, and a diverse package of debugging, display and plotting and mass spectrometer evaluation programs.

Because we view GC/HRMS as the most important new capability of our mass spectrometer/computer work, the requirements of GC/HRMS have guided development of the software system. These requirements include continuous automatic monitoring of instrument performance to avoid wasting time collecting poor or erroneous data. Because we have chosen to approach GC/HRMS with an electrical recording system, as opposed to photographic, we are able to monitor the instrument continuously, both during initial setup and during the course of the GC/HRMS experiment. Major sections of the software and how they interact among one another are summarized below.

1.4 Software Architecture

Figures 2 and 3 show the various software configurations possible within the GC/HRMS system. The data paths and options available in obtaining reference spectra for instrument diagnostics and calibration are illustrated in Figure 2. Successful operation of the mass spectrometer depends on successful setup and verification of the performance of the spectrometer. The various routines outlined in the Figure permit the operator to acquire data and examine it during each stage of

the subsequent reduction. A typical run might be (using the REFRUN command processor for control of the system) to acquire a spectrum (RRØMOD), reduce the scan data to peak times and areas (RRØRED) while saving the raw data on magnetic tape for future reference. Time to mass conversion and display to the operator (on a CRT display) are automatic and provide both the results and diagnostics (RRØREP). These data may be examined further, e.g., via spectrum bar plots, peak profile examination.

The data paths and options available to the operator for collection of high resolution mass spectral data (whether for single samples or for GC/HRMS) are summarized in Figure 3. The various processors summarized in the Figure have a simply-stated but computationally difficult task; to acquire and quickly reduce a spectrum to masses, elemental compositions and intensities. The task is completely automatic. It is based on information about the particular setup of the mass spectrometer (scan time, duration, reference ions, etc.), determined during the reference ions, etc.), determined during the above procedures. Again, raw data are saved for future perusal and the operator can examine the data at any stage during data reduction. Diagnostic and instrument performance information are available after each scan to monitor continuously the status of the mass spectrometer. In normal use, the process runs without interference. Completely reduced data, for the normal spectrum, are available within 2-3 seconds after completion of the scan. This is so much better turn-around than the previous ACME system that it has opened new possibilities for data acquisition and reduction. For example, in cyclic mode, where spectra are acquired repetitively, the computer system can easily keep up with the mass spectrometer. Diagnostics can point out instrument problems before sample is wasted in another run.

1.5 System Philosophy

The underlying philosophy governing the design and development of the software system is dominated by three considerations. First, the operator-data system interface must be flexible enough to meet the changing and often times novel demands required by the experimental nature of the GC/HRMS procedures. While predefined operational sequences are essential for production processing, such sequences must not be so rigidly defined that deviations cannot be made to accommodate experimental modes of instrument operation. Second, the system must maintain its integrity under severe environmental conditions. Unforeseen and often uncontrollable conditions can cause catastrophic hardware failure. The filing system must be made immune to contamination by such occurrences. Third, the overall system structure must be amenable to organized software growth and expansion. It must be easy to add facilities and to implement and evaluate experimental algorithms and heuristics.

1.6 System Flexibility

Due to the dual role of the system as both a production instrument (at this point HRMS) and an experimental instrument, (GC/HRMS and metastable work), the operator-data system interface must provide both a convenient means of executing often utilized operational sequences as well as a flexible means of exploring sequences amenable to the experimental work. Towards this end each major system program consists of a resident command driven interpreter. This interpreter accepts a two letter keyword command from the operator and then invokes an overlaid semantic routine. If an unknown command is entered by the operator, facilities exist for refreshing the operator's memory of which commands are appropriate under the circumstances. Semantic routines may interrogate the operator directly or may utilize default information contained within a disk file. Such a simple structure provides for easy expansion of system facilities while also providing for explicit control of the sequence of operations by the operator.

In addition to the control structure within the PDP 11/45, facilities also exist for controlling the PDP 11/20 directly. Each process which can be loaded into the PDP 11/20 has a finite number of distinct states. Operator commands exist to cause transitions between each of these states. Thus, it is impossible for the PDP 11/20 processor to get 'stuck' waiting for an event which will never occur.

1.7 System Integrity

The minute quantities of certain samples which have been submitted for analysis prohibit the re-running of any experiments associated with these samples. The system operates in a somewhat hostile environment. The physical laboratory environment dictates that the computer system be located in close proximity to the GC/MS instrument. The instrument can cause severe electromagnetic disturbances (sparks within the source, high voltage shut down, etc.) which can bring down either the entire data system or portions of the system. Static electric discharges from the operator through the system console have also resulted in catastrophic consequences for the data system. These occurrences are quite unpredictable from the software point of view and are difficult to alleviate in the physical environment. Therefore, the software must file data as soon as it is acquired in order that in the event of system failure any data gathered up to that point is maintained intact. It is for this reason that the thresholded data is logged directly onto magnetic tape by the PDP 11/20 processor. It is for this reason also that the reduced data filing mechanism insures that the last data block of a scan is actually written out onto the disk and not left in a DOS system buffer.

Some of these topics are amplified in the subsequent sections where several features of the system are described in more detail.

1.8 Operating System

[Between the time this section was drafted and finished, the conversion to the DOS 9 operating system was made.]

DOS version 8 is the operating system currently in use. However, we have converted the software to DOS 9 and are awaiting the installation of the IMS disk system. The major mandate for this conversion is the vastly improved overlay system offered by the new operating system. Overlaid files are maintained as a single, contiguous file on disk as opposed to the DOS 8 method of maintaining a separate linked file for each overlay. The DOS 8 strategy demands that a linked file be opened, read, and closed for each overlay load. DOS 9 allows an overlay to be loaded with a single disk read. Also the DOS 9 overlay facility provides for tree structuring process which was completely absent from DOS 8. Considering that the current version of the system has 17 overlays the importance of efficient overlay loading is obvious. In addition to these factors, DOS 9 provides us with batch processing facilities which make it much easier to do system generation, archive data, etc. The decision to use DOS 9 was a major factor in switching from the System Industries, non-DEC compatible drives to the IMS fully software compatible drives.

1.9 GC/HRMS Software System

On top of DOS the GC/HRMS system has been constructed. This system has both real time as well as non-real time facilities. The real time facilities include: 1) the acquisition and reduction of a reference run to calibrate the instrument and 2) the acquisition and reduction of a sample run. Both processes must install the acquisition routines into the PDP 11/20.

The non-real time facilities include the ability to re-examine reduced data files, to re-run the reduction processes from the back-up medium, to communicate to the PDP 10 or other processors the results of composition matching for use in other processing, for example, mass spectra for MOLION, PLANNER or INTSUM (see Part 2, below).

1.10 General Data Acquisition Procedure

The actual data acquisition procedure is accomplished in two steps. The first step involves calibrating the instrument for the particular set up of the MAT 711 and the second step involves the actual analysis of the sample.

The program REFRUN provides the calibration facilities for the instrument. The operator runs the program and informs the system of the sampling rate, the direction of the mass scan, routing of displays. The final tweaking of the MAT 711 is performed and a scan command is performed. The PDP 11/45

commands the previously loaded process in the PDP 11/20 to start the scan. The 11/20 checks that the MAT 711 interface is set up properly and then initials the scan. When the scan completes the PDP 11/20 is signaled through the mass spec control interface and it compiles a spectrum trailer which it tacks onto the end of the peak profile data it is logging and sending to the 11/45. As the 11/20 feeds these data into the 45, it is crunched down into time intensity pairs. When the spectrum trailer comes through the 11/45 invokes the mass computation algorithm which attempts to locate the prominent landmarks in the PFK spectrum. If this process is successful, a display is produced showing the model peak profile, the resolution as a function of mass, and the projection errors for calibration of the mass/time curve of the instrument. In addition, signal and noise information is displayed. If this process is repeatable (in the sense of taking 2 or 3 scans which yield essentially the same results) and performance at a sufficiently high level, then the reduced data is filed for later use by the sample analysis routines.

The program SAMRUN provides the sample analysis facilities. It is run after a successful REFRUN. The information about the time to mass conversion from the preceding refrun is used to perform the time to mass conversion for the sample spectra. The rapid, automatic nature of this procedure was mentioned above.

1.11 Filing Systems

The filing system can be roughly divided into three components. First, when a spectrum is acquired from the MAT 711 it is thresholded and background removal is performed to produce what is called peak profile data. This data is logged immediately onto magnetic tape by the PDP 11/20. Second, as data is being reduced into mass amplitude pairs by REFRUN or SAMRUN it is filed onto disk for easy retrieval for later examination. The format of the reduced files for SAMRUN and REFRUN is slightly different due to the fact that a refrun contains only one spectrum while a sample may have any number of spectra. Third, the system maintains files which contain control information, spooled hardcopy information and composition output to be transfered to the PDP 10 for further analysis.

1.12 Buffering

Buffering is a central issue in the system. Due to the uneven distribution of data, high data rates, slack periods, it is desirable to provide a large amount of buffering between the instrument itself and the reduction processes. It is the case that data from one spectrum can be reduced while another spectrum is being acquired. Currently the PDP 11/20 has sufficient buffer capacity to hold almost a complete spectrum of a sample. The 11/45 will soon have the capability to run the peak profile to intensity-time reduction process concurrently with the time

intensity to mass amplitude conversion process or the display processes. Such concurrence is another side effect of the operation under DOS 9 due to the ability to tree structure overlays.

This software system permits a great deal of flexibility in operation of the system. Where time between scans permits (a few seconds) the HRMS data can be reduced completely to accurate masses and intensities, and feedback provided to the operator on the quality of the scan. This output is the data used to determine the quality of the spectral data. One can disregard scans which are poor and know when one is of high quality. Alternatively (and in addition to the above mode), data (peak profiles and intensities) can be spooled onto tape for later processing. The operator can choose to print out results immediately for critical samples, or defer final output until later while additional data are being collected. An archival system provides the facility for storing and retrieving old spectral data for review or reanalysis.

1.13 Present Status and Performance Tests

As mentioned previously, the system is operational now and is in routine use for HRMS and in experimental use for GC/HRMS. New developments (see below) and routine use are proceeding in concert. We maintain the previous version of the programs for routine use while work continues on a separate version to add improvements, remove program "bugs" and so forth. We are rapidly proceeding toward a system which is relatively "crash proof" in spite of what the mass spectrometer might do and in spite of abnormalities which might appear in the data. Such a system is critical to ensure the integrity of data collected during a long GC/HRMS run.

We have been running many performance tests on the GC/HRMS system to determine what problems arise when the mass spectrometer and computer system are pushed to their utmost, and to determine the sensitivity in terms of sample size of the GC/MS combination. In several instances, additional programming had to be done to cope with the demands of GC/HRMS. These additions are largely complete now (e.g., allocation of extra memory and disk space for REFRUN when the GC is at high temperature and considerable numbers of ions from GC column bleed are present in addition to the internal mass standard, perfluorokerosene) and we have turned our attention to measuring performance samples.

In performing sensitivity tests we can always make the system look good by choosing samples which have characteristics such that excellent mass spectra can be obtained on minute amounts of material. We have not done this. We have chosen samples which are representative of the types of material that are the focus of our current chemical research interest. For the simpler case of fatty acid methyl esters, we can obtain, in 8-10 sec. per decade in mass scans, HRMS displaying ions over a

dynamic range of 100:1 with about 1 microgram of material per component. For the harder case of free sterols, such as cholesterol, 2 micrograms per component yields the same performance (such sterols have many more ions over a wider mass range, thus requiring more material for the same dynamic range of ion intensities). We do not claim that this sensitivity is the ultimate achievable. It is certainly sufficient for many of our problems. It will be insufficient for those problems where there are components over a wide range of concentration. Because the GC column cannot be overloaded too severely for the major components it is difficult to increase greatly the concentration of the minor components. Additional chemical or physical separations can solve some problems of this type. But with this sensitivity we can get much useful work done as we progress with improvements to our techniques (see Future Developments).

Current applications of the mass spectrometer/computer system to biomedical structure elucidation problems are summarized in Part 3 of this report.

1.14 Future Developments

The second year of our grant has several specific goals for the mass spectrometer/computer system. These goals, outlined below, will improve the performance and reliability of the current system to ensure the integrity of results on precious samples. This year will also be taken up with implementing the other hardware and software items mentioned in our original proposal (metastable ion work, MS-9 hook up, multiplet detection and resolution) now that the primary goal of GC/HRMS is well in hand.

1.14.1 Hardware

The following are the important goals in hardware development, necessary to fulfill our research objectives.

A) Installation and testing of the hardware for control of the mass spectrometer for semi-automatic acquisition of data on metastable ions. This hardware, including circuitry to interface the metastable scanning system of the Varian MAT 711 to the 11/45 system, and a high precision A/D converter, will permit semi-automatic detection and analysis of metastable ions which relate a "daughter" fragment ion to its progenitors, "parents". It will allow us to explore the inverse relationship, determination of all daughter ions from a given parent ion by simultaneous variation of two of the three fields in the instrument, accelerating voltage, electrostatic analyzer voltage and magnetic field. The feasibility of this technique has apparently been demonstrated by Lacey and McDonald in Australia.

B) Reconnection of the MS-9 to the new 11/45 data system.

There is a relatively minor amount of work which must be done to enable us to acquire data from the MS-9 under the new instrument control structure of the 11/45 computer system. This will enable us to divert routine samples to the MS-9 for analysis and permit us to devote more time on the 711 to the more difficult task of GC/HRMS.

C) Connection of a plotter to the computer system. An existing Cal-Comp plotter will be connected to the system so that hard copy of graphical (e.g., mass spectra, instrument performance curves) output can be obtained. Presently only CRT output of this information is available.

1.14.2 Software

With the hardware largely installed and functioning, the software (the actual programs which acquire, manipulate, reduce and output the mass spectral data), requires the greatest attention in the coming year. There are several steps to be taken which will improve the capabilities of the GC/MS system in general, GC/HRMS in particular. These are outlined below.

A) Improve data reduction facilities for scanning at lower resolving powers. HRMS data are usually collected at a resolving power sufficient to separate many (but never all) of the possible multiplets of ions possessing the same nominal atomic mass but differing in elemental composition. However, for maximum sensitivity, with relatively little degradation in data quality, one would like to run the mass spectrometer at lower resolving powers. This increases the the likelihood of overlapping peaks which are viewed as single peaks according to our present data acquisition system. Our proposal discussed ways to use both mathematical routines and chemical intelligence to help solve this problem, thus providing effectively higher resolution via data processing, at high sensitivity. We are just implementing the first phase of this approach, to resolve quickly those overlapping peaks whose profiles are well-defined. We view these developments as essential to the success of GC/HRMS because of the improved sensitivity.

B) Better inter-computer communication. We are currently implementing better inter-machine communication between the 11/20 and 11/45 computers (see Fig. 1). This will improve the reliability of the system by allowing "clean" (i.e., restartable) recovery from error conditions in the mass spectrometer, in the input data stream or during data reduction between scans.

C) Implement a GC/LRMS system. Because of our focus on GC/HRMS, the ability to handle efficiently GC/LRMS data has been neglected. We will remedy this situation in year two because some of our problems do not require HRMS data and rapid presentation of LRMS data to the chemist in graphical form will be very useful. The LRMS system will be relatively simple to implement because almost all of the same data acquisition and reduction programs written for HRMS data can be utilized.

D) Software for metastable ion analysis. Routines must be written to enable facile calibration of the mass spectrometer operating in metastable ion mode, and to allow subsequent reduction of data. Existing control and data acquisition software will be used for initiating the metastable ion mode.

1.14.3 Summary

As the above hardware and software improvements are being made we will continue evaluation of the GC/HRMS system in parallel with its actual application to real problems. GC/HRMS is a relatively new and difficult technique for routine application. In order to use it effectively, we will have to exert some effort toward determining and optimizing the performance of the many elements of the system, the GC, the MS, and the computer hardware and software.

2 PART 2: DEVELOPING PERFORMANCE AND THEORY FORMATION PROGRAMS

TO ASSIST IN BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS

2.1 Introduction

The Heuristic DENDRAL computer programs assist with structure elucidation problems by helping interpret mass spectra and helping generate structures that are consistent with the interpretations. The Meta-DENDRAL programs assist with rule formation problems in cases where the rules of mass spectrometry are not known.

In this section we describe our progress on the computer programs. Generally speaking, it has been a productive year because of the interactive computing environment provided by the SUMEX facility. Not only are we able to develop programs much more rapidly than in a batch environment, but we are able to make the programs themselves highly interactive, and thus more useful.

All programs have been transfered to the SUMEX machine and most have been considerably improved from their previous versions. The CONGEN and PLANNER programs, in particular, have been improved substantially because these two were thought to offer the most to scientists with structure elucidation problems. Two new programs were developed in this period: CLEANUP and MOLION. The CLEANUP program helps separate the mass spectra of individual components from a GC/MS analysis, and eliminates the background due to GC column "bleed". The MOLION program determines the mass and empirical formula of the whole molecule from its mass spectrum, without prior knowledge of any of the features of the molecule. Both of these new programs solve major

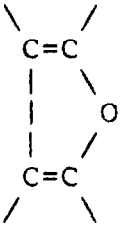
problems that we had previously assumed were already solved before a scientist used the DENDRAL programs.

2.2 CONGEN.

The CONGEN[48,53] program represents a significant extension of a program which has developed over the last several years, the cyclic structure generator[40,41]. The purpose of CONGEN is to assist the chemist in determining the chemical structure of an unknown compound by 1) allowing him to specify certain types of structural information about the compound which he has determined from any source (e.g., spectroscopy, chemical degradation, method of isolation, etc.) and 2) generating an exhaustive and non-redundant list of structures that are consistent with the information. The generation is a stepwise process, and the program allows interaction at every stage; based upon partial results the chemist may be reminded of additional information which he can specify, thus limiting further the number of final structures.

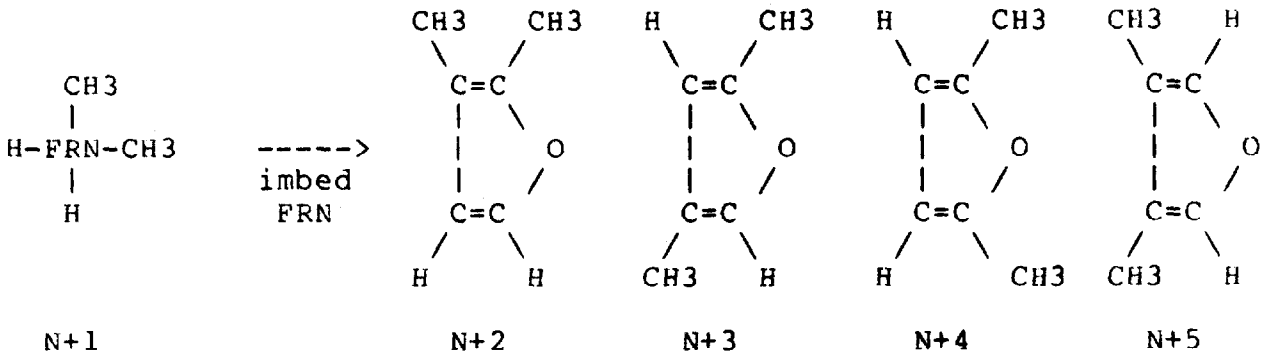
CONGEN fits with the other DENDRAL programs as a "backstop" solution to structure elucidation problems. If the mass spectrum of an unknown compound is available, then CLEANUP and MOLION could be used, but if the general class of the compound is not known, PLANNER has no starting point from which to work. In such cases, structural information can be extracted manually from the spectrum and given to CONGEN for analysis. Because CONGEN makes no assumptions about the source of this information, other spectroscopic or chemical techniques may be used to supply supplemental data.

At the heart of CONGEN are two algorithms whose accuracy has been mathematically proven and whose computer implementation has been well tested. The structure generation algorithm[31,37,40,41] is designed to determine all topologically unique ways of assembling a given set of atoms, each with an associated valence, into molecular structures. The atoms may be chemical atoms with standard chemical valences, or they may be names representing molecular fragments ("superatoms") of any desired complexity, where the valence corresponds to the total number of bonding sites available within the superatom. For example, in a compound known to contain a furan ring, the quadrivalent superatom FRN might be defined, which has the structure N. Here, the bonds with



N

unspecified termini represent available bonds to hydrogen or other atoms or superatoms. Because the structure generation algorithm can produce only structures in which the superatoms appear as single atoms (we refer to these as intermediate structures), a second procedure, the imbedding algorithm[48,53] is needed to expand the superatoms to their full chemical identities. For example, N+1 is a simple intermediate structure



which might be produced by the structure generator. The imbedding of FRN yields four final structures, N+2-N+5. The output of the imbedder is exhaustive and, in a limited technical sense, free from duplication. But when a list of intermediate structures undergoes imbedding, duplicates can arise. Thus the imbedder is also equipped to post-test such lists for duplicates and retain only unique structures.

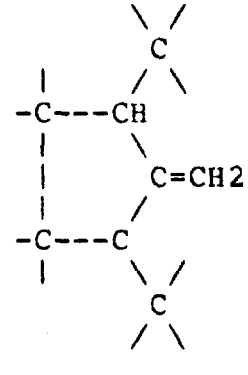
These two routines give the chemist the ability to construct structures from a given set of molecular "building blocks" which may be atoms or larger fragments. By itself, this capacity is of limited utility because the number of final structures can be overwhelming in many cases. Usually, the chemist has additional information (if only some general rules about chemical stability, which the program has no concept of) that can be used to limit the number of structural possibilities. For example, he may know that because of a compound's stability, it cannot contain a peroxide linkage (O-O) and thus the programs need not consider such structures when there are two or more oxygens in the "building block" list. During the past year, a substantial amount of effort has been devoted to modifying these two basic procedures, particularly the structure generation algorithm, to accept a variety of other structural information (constraints), using it as efficiently as possible to prune the list of structural possibilities.

Specifically, there are six types of constraints that we have implemented. GOODLIST and BADLIST are used respectively to require and forbid the presence of user-specified substructures in intermediate or final structures. The peroxide constraint above could be specified on BADLIST, for example. On GOODLIST are placed desired substructures which cannot be entered as superatoms either because their number is uncertain (each GOODLIST entry has an associated minimum and maximum number of occurrences), or because they may share atoms with other superatoms (the "building blocks" must be mutually disjoint

units) or because they are not sufficiently precise: The substructure X=X-X=X, where X represents any non-H atom, may be used on GOODLIST to require a conjugated system, but because it is not composed of specific chemical atoms, it cannot be used as a superatom. Two other constraint types, GOODRINGS and BADRINGS, are similarly used to require or forbid the presence of user-specified ring sizes in final structures. The PROTON constraint, especially useful in specifying information from proton NMR spectra, is used to specify desired numbers of protons in specific chemical environments, while the ISOPRENE constraint filters out final structures which do not obey the isoprene rule, an important rule in natural-products chemistry.

The remainder of the work on CONGEN during the year has gone into the human engineering needed to develop an easily used, interactive "front end" to these routines. In this category we include EDITSTRUC, an interactive structure editor, DRAW, a teletype-oriented structure display program and the CONGEN "executive" program which ties the individual pieces together and aids the user with various tasks, such as defining superatoms and substructures, creating lists of constraints and "building blocks" and saving and restoring superatoms, constraints and structures from secondary storage (disc). The resulting system, for which comprehensive user-level documentation has been prepared, is running on the SUMEX computing facility and is available nationwide over the TYMNET and ARPANET networks. Several chemists are currently using CONGEN to assist them in structure identification problems.

Although it is still an unsolved structure, a recent case for which CONGEN was used exemplifies the current capabilities of the program. The compound is a marine sesquiterpene of empirical formula C15H24 (with unsaturation, that is number of rings plus multiple bonds, equal to four double-bond equivalents) and there is quite a bit of spectroscopic and chemical data available. These data indicate the presence of a superatom (arbitrarily) called Z, corresponding to structure N+6, along with an isopropyl group (superatom IP) and one



N+6

additional methyl group (superatom ME). These superatoms, together with three more carbon atoms and twelve more protons (or, equivalently, two additional degrees of unsaturation) constitute the "building block" list. There are three parts to

the problem because the superatom Z can have zero, one or two internal bonds created by joining zero, one or two pairs of free valences within Z. Each sub-case represents a separate problem for CONGEN. If no additional constraints are used, the structure generation algorithm constructs 487 intermediate structures for the three cases together, and it can be estimated that the imbedding of Z, IP and ME would produce many thousands of final structures.

For this compound, however, additional information is available. The BADRINGS feature is used to specify that no three-membered rings (which would produce characteristic absorptions in the proton NMR spectrum) may be produced. There is evidence that no multiple bonds other than the exocyclic double bond in Z exist in the molecule, so GOODRINGS is used to require exactly one "two-membered" ring in final structures (CONGEN views multiple bonds as small rings). The fact that there are no methyl groups beyond those in IP and ME is expressed through BADLIST. The final two constraints, that IP must be connected to a methine carbon and ME to a quaternary one, can be entered on either GOODLIST or BADLIST (for the latter, the forbidden environment of IP and ME are specified), and because BADLIST is implemented more efficiently, it is preferred. With these constraints, the structure generation algorithm obtains only 16 intermediate structures for the three sub-cases, in roughly one-third the time required for the unconstrained generation. The imbedding of Z, IP and ME, using the same constraints, give 179 final structures in all.

There is also off-resonance decoupled C-13 NMR data indicating that the compound possesses three methyl groups, five each of methylene and methine carbons and two quaternary carbons. One would expect this data to substantially limit the final structures, but in fact none are eliminated by this GOODLIST constraint. Thus, the "degree sequence" of the carbons is implied by the other constraints, a rather unexpected result.

Additional work on this problem has been carried out using GOODLIST constraints based on analogy with other compounds isolated from the same source and identified. We currently have a set of four structures which represent the most plausible candidates, but a rigorous identification of the compound has not yet been made.

The CONGEN program has just reached the borderline of practical "real world" problems a chemist is likely to face. Although there are some practical cases in which CONGEN has been and is being used, the probability is high that a typical new case, as it is naturally input by the chemist, will either run for an excessively long period of time or will use up all available core storage, or both. There are two facets to this problem. On one hand, the programming has been done primarily in INTERLISP, a language in which the development of complex programs can take place with relative ease. Though this language has assisted in the rapid progress of CONGEN, it is quite inefficient in both execution time and core requirements. We

estimate that the recoding of the most time-consuming portions of the program would speed these portions by a factor of 50-100 and would significantly reduce the demand upon computer memory. On the other hand, there are many cases in which structures are not tested against some of the constraints until after the structures have been generated. In highly constrained cases, this post-testing can be time-consuming; large numbers of structures are generated only to be discarded later. A great deal of research remains to be done in the "intelligent" use of constraints within the program so it can distinguish, early in the analysis, those logical sub-cases which will produce no acceptable structures.

During the next year we plan to address these two problems directly. CONGEN already contains portions in SAIL and FORTRAN, languages much more efficient than INTERLISP, and we plan to recode some of the more time-consuming and less developmental portions of CONGEN into one or the other of these languages. Research, ranging from the discovery of new programming "tricks" to the improvement of the basic mathematics underlying CONGEN, will proceed in the direction of efficient utilization of constraints. Paralleling this will be the continuing development of the "visible" portion CONGEN (the "executive") which, with the guidance of chemist collaborators, will be made increasingly flexible and easy to use. All of these developments will be guided by our desire to make CONGEN a responsive and practical resource for the chemist.

2.3 PLANNER

The DENDRAL PLANNER program [28,33] is designed to analyze the mass spectrum of a compound or of a mixture of related compounds. Because there is no ab initio way of relating a mass spectrum of a complex organic molecule to the structure of that molecule, PLANNER requires fragmentation rules for the class of compounds to which the unknown belongs. This is its major limitation.

Applications and limitations of PLANNER have been discussed extensively. [28,53] The program is very powerful in instances where mass spectrometry rules are strong (i.e., general, with few exceptions). In instances where rules are weak or nonexistent, additional work on known structures and spectra may yield useful rules to make PLANNER applicable (see INTSUM and RULEGEN, below). One unique feature of PLANNER is its ability to analyze the spectra of mixtures in a systematic and thorough way. Thus, it can be applied to spectra obtained as mixtures when GC/MS data are unavailable or impossible to obtain.

The power of the PLANNER has been substantially increased by including the MOLION program (discussed below) as a subroutine for computing the list of plausible molecular ions. Since this subprogram does not depend on knowledge of the compound class, the PLANNER no longer needs to have class-specific rules for determining the mass and empirical formula of the unknown molecule.

The major developments to the PLANNER program have been in the so-called "user interface" - the language and prompts typed by the program to the chemist user. Once the program was successfully transferred to the SUMEX machine, including rewriting parts of it, it was possible to make the program truly interactive. It requires three pieces of information as input from the chemist: the high or low resolution mass spectrum, the characteristic skeletal structure for molecules in the specific compound class, and the fragmentation rules for the class. Additional knowledge about the unknown can be used optionally by the program.

The interactive nature of the program has been enhanced by development of "help" facilities. For example, when a person does not know how to respond to one of the program's prompts, he can always get some help by typing a question mark. An annotated typescript of an interactive session of the PLANNER is shown in Appendix 1.

A chemist's interaction with this program has also been simplified by providing the structure definition and drawing facilities of the CONGEN program (discussed above). Thus it is easy now to tell the program the skeletal structure of the molecules in the compound class. The whole package is relatively natural and easy for a chemist.

We have also added save and restart capabilities to this program so that a chemist can avoid redefining the parameters for a class. This is useful when a chemist needs to explain spectra from more than one sample from the same class. This capability also allows cumulation of knowledge of different classes. For this purpose it is still primitive, but it is a necessary first step.

Along with making the program's parameters accessible to chemist-users, we have also been making the parameters more general to increase their power. For example, the parameter CONTROLRULES, which controls the way the program thresholds evidence, can be set by a chemist to apply to specific individual fragmentations (instead of all fragmentations uniformly). This is a useful way of communicating to the program the heuristic that a given fragmentation process is strong enough and reliable enough that only the strongest evidence for this fragmentation needs to be considered at first. (The program can also be instructed to relax this restriction on this fragmentation later.)

2.3.1 Future Plans

Since the PLANNER program has tentatively moved from a research program to a working laboratory tool, we have discovered a number of ways to tailor the program to the needs of mass spectrometrists. These fall into two general categories: making the program easier to use and making the program more powerful.

The interface will be made "smoother" so that chemists will have an easier time with the interactive dialogue. This means both improvements to the language of the dialogue and improvements to the control structure. The latter is necessary to help the program recover from the errors and from user interruptions.

Additional heuristics will be put at the disposal of users. In particular, specific structural features may be thought to be present or absent in the unknown and we want the PLANNER to use this information as easily as the CONGEN program does. Another improvement will be coupling this program with CONGEN.

We are also ready now to increase the scope of the program by making it work with small structural skeletons and fragmentations involving substituents on a skeleton. For example, aromatic acids are probably best described as an aromatic ring with various substituents. Since the aromatic ring itself does not fragment in characteristic ways, all of the breaks must be described as breaks in the substituents. These improvements mean that many more classes of molecules will be amenable to analysis and many extra types of fragmentations can be used in the analysis.

Finally, we are planning to increase the efficiency of the program. We have already made some of the preliminary timing tests so we know where to focus our attention first. Secondarily, we can also reduce system overhead by compiling the program in blocks, which we plan to do.

2.4 CLEANUP

The raw data obtained from GC/LRMS analysis of complex mixtures (e.g. urine samples) consists of a large number (600 to 1000) of mass spectra that result from sampling the GC effluent over an extended fractionation period. Because of the limited separation capability of the GC and contamination by the liquid phase of the GC column most of the spectra obtained are not directly useable. We have developed a program called CLEANUP which takes these raw mass spectral data and removes column bleed and contributions from partially overlapping neighboring elutants.

The CLEANUP program outputs a set of "clean" mass spectra suitable for library matching and/or analysis by other DENDRAL programs. The data reduction factor from the raw data to the cleaned up data is usually an order of magnitude depending on the complexity of the mixture.

The CLEANUP program is directly linked to our MOLION and library search programs. This makes it possible for us to go through the process of data-collection, data-reduction and library search in a smooth automated mode. We obtain as output from this process a line printer listing of possible compounds

for each component found in the mixture. Steps are being taken to extend the system so that components found that are not in the library will be automatically flagged for later reference and subsequent analysis by other DENDRAL programs. We are in the process of writing a manuscript which will describe the procedure of CLEANUP in detail, with examples.

2.5 MOLION

After running a mixture through the GC/MS and CLEANUP, we are left with a collection of more or less pure spectra of unknown compounds. Structure elucidation now begins in earnest. The key elements in problems of structure elucidation are the molecular weight and formula of a compound. Without these absolutely essential data, the structural possibilities are usually too immense to proceed further. Mass spectrometry is frequently used to determine molecular weights and formulae. However, there is no guarantee that the mass spectrum of a compound displays an ion corresponding to the intact molecule. For example, many of the amino acid derivatives in urine samples display no molecular ion. When we are given only the mass spectrum (and for this type of GC/MS analysis a mass spectrum may be all that is available) we must somehow predict likely molecular ion candidates. The new program MOLION [45] performs this task. Given a mass spectrum, it predicts and ranks likely molecular ion candidates independent of the presence or absence of an ion in the spectrum corresponding to the intact molecule.

The MOLION program is written to operate on either low or high resolution mass spectra. It is insensitive to the type of compound analyzed. The program has certain limitations which have been summarized in detail previously.[45] Briefly, the program has difficulties with spectra containing ions from higher molecular weight impurities (thus, the need for CLEANUP, discussed below) and with spectra which contain only low mass ions (representing less than half the weight of the intact molecule). These, of course, are instances where manual interpretation also has most difficulties. The program is currently being modified so that it can cope with spectra of mixtures of compounds.

MOLION is available on SUMEX. A FORTRAN version, initially for low resolution mass spectra, is being written so that the program can be run on smaller computers and exported to others. However, it will continue to be available via SUMEX so that others can access it easily. It is available as a stand-alone, interactive program and also as a part of the PLANNER program.

2.6 Library Search

Over the course of several years, libraries of mass spectral data have been assembled (e.g. GE-network library developed at NIH, the Markey library, the Aldermaston library).

We are presently using spectra from these libraries together with our own compilations to augment our library search facility. Library search provides us with an efficient mechanism for weeding out from a group of spectra those which represent known compounds. Known, commonly occurring components, usually make up more than half of any typical biological sample that we analyze. Clearly, one should spend time on solving the structures of unknown compounds, not rediscovering old ones. The CLEANUP program provides mass spectra which are of sufficient quality to expect that known spectra should be identified relatively easily from such libraries. (Without representative spectra, only total nonsense will result from matching spectra to a library.)

Experience with available libraries has shown that even though these compilations are extensive they are not entirely adequate for analysis of spectra from urine extracts. To cope with this inadequacy we have made an effort to make our library management facility as flexible as possible for updating and modification. As new compounds are identified they are added to our own library compilations for future use.

We are currently investigating the possibility of using a low resolution version of MOLION to enhance the selectivity of the current algorithm. GC retention indices are assigned to the spectra we collect. Some use may be made of these indices in future versions of the program should they be needed.

2.7 MetaDENDRAL Rule Formation Programs

The INTSUM program [34] is in routine, production use to assist in interpretation of the mass spectra of new classes of molecules (see Part 3 for details). When the mass spectrometry rules for a given class of compounds are not known, the INTSUM, RULEGEN and RULEMOD programs can help a chemist formulate those rules. Essentially, these programs categorize the plausible fragmentations for a class of compounds by looking at the mass spectra of several molecules in the class. All molecules are assumed to belong to one class whose skeletal structure must be specified. Also, the mass spectra and the structures of all the molecules must be given to the program.

INTSUM collects evidence for all possible fragmentations (within user-specified constraints) and summarizes the results. For example, a user may be interested in all fragmentations involving one or two bonds, but not three; aromatic rings may be known to be unfragmented ; and the user may be interested only in fragmentations resulting in an ion containing a heteroatom. Under these constraints, the program correlates all peaks in the mass spectra with all possible fragmentations. The summary of results shows the number of molecules in whose spectra there is evidence for each particular fragmentation, along with the total (and average) ion current associated with the fragmentation.

The RULEGEN program attempts to explain the regularities

found by INTSUM in terms of the underlying structural features around the bonds in question that seem to "drive" the fragmentations. For example, INTSUM will notice significant fragmentation of the two different bonds alpha to the carbonyl group in aliphatic ketones. It is left to RULEGEN to discover that these are both instances of the same fundamental alpha-cleavage process that can be predicted any time a bond is alpha to a carbonyl group.

A new development has been the RULEMOD program for modifying and condensing the set of rules produced by INTSUM and RULEGEN together. It looks at the negative evidence associated with each candidate rule in order to select the best ones, then merges rules that seem to explain the same breaks (if possible).

The Meta-DENDRAL programs RULEGEN and RULEMOD have now developed to a point that the programs have rediscovered the mass spectrometry rules for two classes of chemical compounds and substantially aided in the search for explanatory rules for a new family. Chemists are now able to use preliminary versions of these programs profitably. We have used eleven aliphatic amines (and their low resolution spectra) and ten estrogenic steroids (and their high-resolution spectra) as test cases. And we have just run thirteen keto-androstanes as the first test of the new programs on a previously-uncharacterized class of molecules. Although we are attempting to find characteristic rules without necessarily finding explanations for all the data, the final sets of rules are sufficient to explain between one-third and two-thirds of the total data (measured as total ion current).

2.8 Results

2.8.1 Amines

The low resolution spectra of the aliphatic amines are highly ambiguous insofar as many different processes can be generated to explain any one peak. In INTSUM, therefore, we limited the range of interesting processes to those producing a nitrogen-containing ions (high resolution mass spectra of some of these amines reveal that almost all ions contain the nitrogen atom). Processes breaking one or two bonds (with +2 to -2 hydrogen transfers) were considered. The output from INTSUM correlates peaks with breaks that are n bonds away from the nitrogen atom. The output is voluminous because there is some low resolution peak corresponding to almost every one- and two-bond fragmentation for these amines. Also the INTSUM output shows almost no consistent regularities because it looks for regularities in terms of a fixed skeleton which, in this case, is small (-N-).

RULEGEN reduces the break information to eleven rules -- a rule mentioning a bond environment (a subgraph) and a set of bonds in that environment that can be expected to break, with

hydrogen transfers. RULEGEN is attempting to explain the regularities noticed by INTSUM in terms of the bond environments that "drive" the common processes. Each rule is well-supported by the data, but there is still some redundancy in the rules. That is, RULEGEN selects individual rules on the basis of evidential strength without considering the set of rules as a whole.

A new program, RULEMOD, reduces the output of RULEGEN still farther by selecting the "best", or most comprehensive, of the rules. The guiding principle here is that a mass spectral peak only needs to be explained once. (There are cases where this is known to be false -- peaks get contributions from several different processes -- but because there is no way to say in advance which peaks deserve more than one explanation, we let economy of rules guide the selection.) RULEMOD first selects the rule that explains the most peaks (1) , then removes those peaks from the evidence supporting the other rules. Then, recursively, the program selects the next best, and so on until the remaining rules have no evidence that is not already explained. For the amines, this step reduces the eleven rules to five.

RULEMOD then considers the possibility of refining the rule set still more by "merging" rules that are very similar. If the subgraphs defined in rules R1 and R2 differ by very little -- i.e., almost every time R1 applies R2 will apply, and vice versa -- then RULEMOD looks for a slightly more general form of the subgraph that will include both R1 and R2 and that will not generate any new negative evidence. This is the first point at which negative instances are considered because, up to here, the number of rules is too large to graph-match against all molecules. For amines, this refinement step rejected all the mergings considered. So the final set of rules for the aliphatic amines, explaining two-thirds of the total ionization of all the spectra, were the five selected previously: alpha cleavage and four two-bond fragmentations (alpha cleavage + cleavage next to N; alpha cleavage + beta cleavage; beta cleavage + cleavage next to N; breaking off two alkyl fragments to keep an arbitrarily large nitrogen-containing fragment).

2.8.2 Estrogens

The high resolution spectra for estrogens made the INTSUM output more specific than for amines, but it was still voluminous. Evidence was gathered for all processes in which either 2 or 3 bonds were broken in the estrogen skeleton (without breaking the aromatic ring, without breaking two bonds to the same carbon, etc.). In each molecule four to ten fragmentation processes showed strong evidence in the corresponding spectrum.

---

(1) "most" can be defined with respect to number of peaks, percent ionization, or almost any other measure -- we used number of peaks here, giving a higher weight to peaks that can be explained only by the rule in question.

RULEGEN produced 26 rules explaining the INTSUM data (INTSUM looks for regularities and does not try to "cover" all of the data points). RULEMOD selected 15 of those 26 that could explain all of the peaks explained by the 26 rules. Then, in the merging step, RULEMOD merged five pairs of rules together to produce 10 rules explaining one-third of the total ionization in the estrogen spectra. These included descriptions of the well-known breaks B,C,E,F (but not D), cleavage of ring-B next to the ring B-C fusion, cleavage of ring-C next to the ring C-D fusion, and four other processes involving three bonds each. These rules would be considered "good" (generally useful) rules by a mass spectroscopist. Thus, the step of interpreting the INTSUM output and removing ambiguities is now amenable to automation.

2.8.3 Keto-androstanes

The keto-androstanes have not been previously studied and are not as well-behaved as the estrogens. The INTSUM output from the high resolution spectra shows no consistent regularities involving one or two bonds in the environment of the keto group. There is evidence for many fragmentations, but none of it overwhelmingly recommends any fragmentation as being universal or even common among all the members of the class.

RULEGEN first looked at subgraphs large enough to encompass alpha-cleavage (up to 2 atoms away from the bonds broken), but still found no regularities for the whole set of data. We enlarged the bond environment to encompass beta cleavage (up to 3 atoms away from the bonds broken) and tried again. Over 80 rules were produced as possible explanations of the one and two-step fragmentations noticed by INTSUM. About one-third of the total data (36%) were explained by these rules. There was considerable overlap in the rules, because RULEGEN's local view of any rule prevents it from noticing that very similar descriptions of bond environments were produced at different points in its search. RULEMOD found that 20 of these rules could explain all of the data that the 82 rules explained. None of the proposed mergings were acceptable to RULEMOD because it allows a result of merging to have no additional negative evidence that the merged rules alone did not have. (This is a very conservative strategy and needs more study.) This set of 20 rules could be reduced to 13 rules by throwing away the 7 rules whose evaluation scores are below zero (indicating that these rules had more negative evidence than positive evidence). This would reduce the total amount of data explained from 36% to 27% but increases the human readability of the rule set. Such thresholding will be added to the program as an option. We are currently examining additional keto-androstanes and will describe this work shortly.

2.8.4 Future work

2.8.5 New Experiments